Research Paper Title:

PT-AML: Machine learning framework to identified personalized treatments for acute myeloid leukemia

Journal: ***TBD***

Authors: Raghvendra Mall[1#], Siddhi Jani[2], ...

Keywords: []

SC: Deep Learning; Machine Learning; AML; Drug Sensitivity; Precision Medicine; Mutations; Cell State; Mechanism of Action

Abstract: **[Checklist:** Statements are factually accurate; provides a concise summary that is interesting and attracts attention; provides key background, purpose of article, and key findings from the study**]**

ABSTRACT EXAMPLE:

First line: State the topic; should be related to the title

Resistance to cell death is a leading hallmark of cancer. Therapies aimed at activating cell death are therefore of high interest to improve treatment and understanding the induction of cell death during cancers is a key strategy to identify therapies.

Second and third sentences: Expand on the critical background information relating to the topic, including why it is a relevant

Fourth sentence and beyond: These sentences should mirror the subsections from your results outline

.

Last sentence: These results demonstrate…

SC:

Research Article Outline:

PT-AML: Machine learning framework to identified personalize treatments for acute myeloid leukemia

[Checklist: Statements are factually accurate, key references from Kanneganti lab and Others ref lists are included**]**

1. **Introduction** **[Checklist:** limit to 3 paragraphs; describe the key concepts; provide relevant background information in the context of our lab's interests; attract the readers attention and inform about the purpose of the article and what you aim to achieve**]**

SC:

2.

3.

4.

5.

6. Discussion

Figures

[Checklist: concise title and legend; include labels and enough detail to be understood without the text; fonts and colors follow lab template, and the formatting is correct; no image duplications**]**

Figure Checklist

First authors and coauthors are responsible for ensuring that: 1) data are reproducible and there are no image duplications; 2) data have been independently verified by 2 people other than the first author (provide their names here).

See the example figure on the next page for clarifications. Please carefully check each item and sign off on the completion of this checklist before bringing figures to Thiru or Rebecca. Finalized figures must be approved by Thiru before you begin writing the manuscript.

Many of these points also apply for review figures. Particularly, for review figure color scheme ideas, refer to the colors used in this poster from InvivoGen: https://www.invivogen.com/sites/default/files/invivogen/resources/documents/2016-poster_tlr-nlr-invivogen_0.pdf

[1] Figure # is at the bottom right of each figure in Arial font, size 14, bold (ex: **Figure 1**)

[2] Figure title is at the top of the page in Arial font, size 11, bold (ex: **Figure title**)

[3] Figure panels are denoted by bold, uppercase letters in Arial font, size 14 (ex: **A**)

[4] All text within the figure panels is in Arial font, size 11, not bold (ex: Media)

[5] Western blot images include labels for the lanes along the top of the image; other labels are also at the top

[6] Lane and border thickness are set to 1 pt with dotted lines on western blots set to 0.5 pt

[7] No shadows or shadow lines are present

[8] Terminology and labeling are used consistently throughout all figures (ex: *Casp11*$^{-/-}$)

[9] Colors should be consistent throughout all figures (wild type, black; knock-out, red)

[10] Red and green are not included on the same graph (due to red-green colorblindness)

[11] All microscopy images include the scale bar, which is defined in the figure legend

[12] Arrows can be used in microscopy images to callout cell death or other noteworthy staining; these arrows should be triangles (ex:)

[13] All blots and gels have molecular weights (or base pair size for DNA/RNA) for all proteins or DNA/RNA species indicated

[14] All blots must have an accompanying raw blot file (see example in Figure 2)

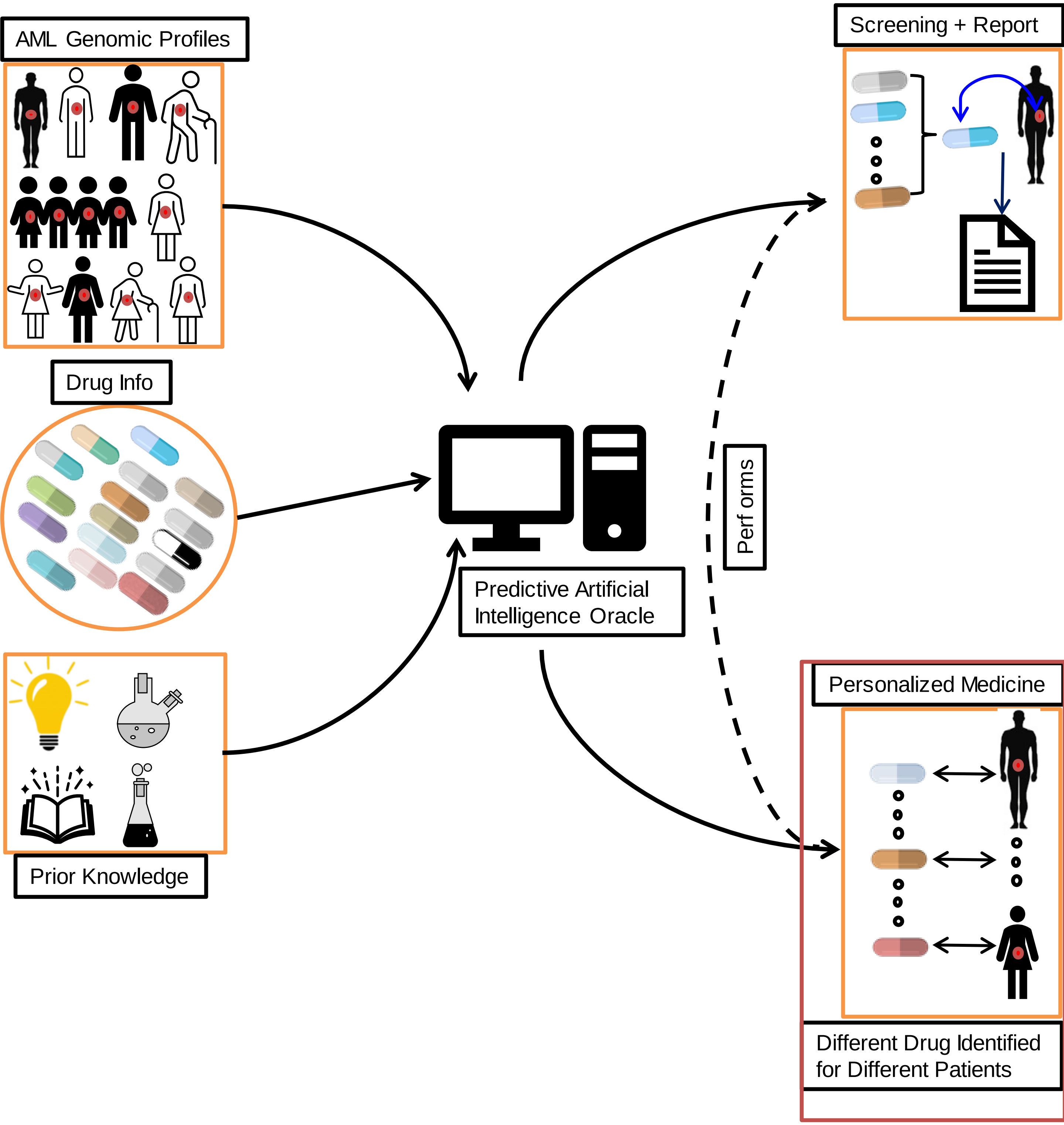
[15] **There are no image duplications** (no duplicated blots, microscopy images, etc.) and data are reproducible and have been validated by 2 additional people as listed above

I confirm that my figures meet the above criteria, along with any other journal-specific criteria.

Name Date

PT-AML: Machine learning framework to identified personalized treatments for acute myeloid leukemia

Graphical Abstract



AML Genomic Profiles

Drug Info

Prior Knowledge

Predictive Artificial Intelligence Oracle

Perf orms

Screening + Report

Personalized Medicine

Different Drug Identified for Different Patients

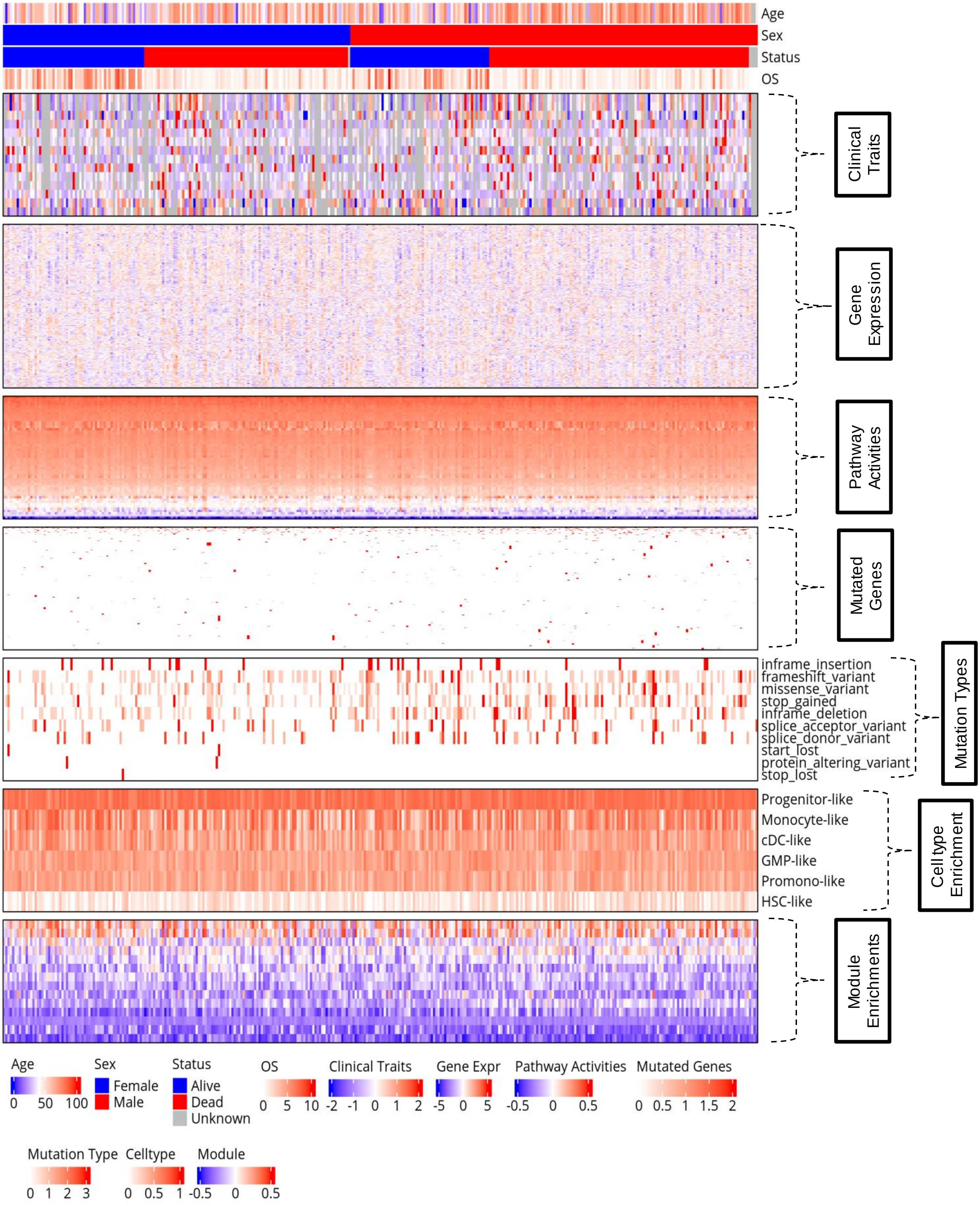

Age
Sex
Status
OS
Clinical Traits
Gene Expression
Pathway Activities
Mutated Genes
inframe_insertion
frameshift_variant
missense_variant
stop_gained
inframe_deletion
splice_acceptor_variant
splice_donor_variant
start_lost
protein_altering_variant
stop_lost
Mutation Types
Progenitor-like
Monocyte-like
cDC-like
GMP-like
Promono-like
HSC-like
Cell type Enrichment
Module Enrichments
Age
0 50 100
Sex
Female
Male
Status
Alive
Dead
Unknown
OS
0 5 10
Clinical Traits
-2 -1 0 1 2
Gene Expr
-5 0 5
Pathway Activities
-0.5 0 0.5
Mutated Genes
0 0.5 1 1.5 2
Mutation Type
0 1 2 3
Celltype
0 0.5 1
Module
-0.5 0 0.5

Figure 1

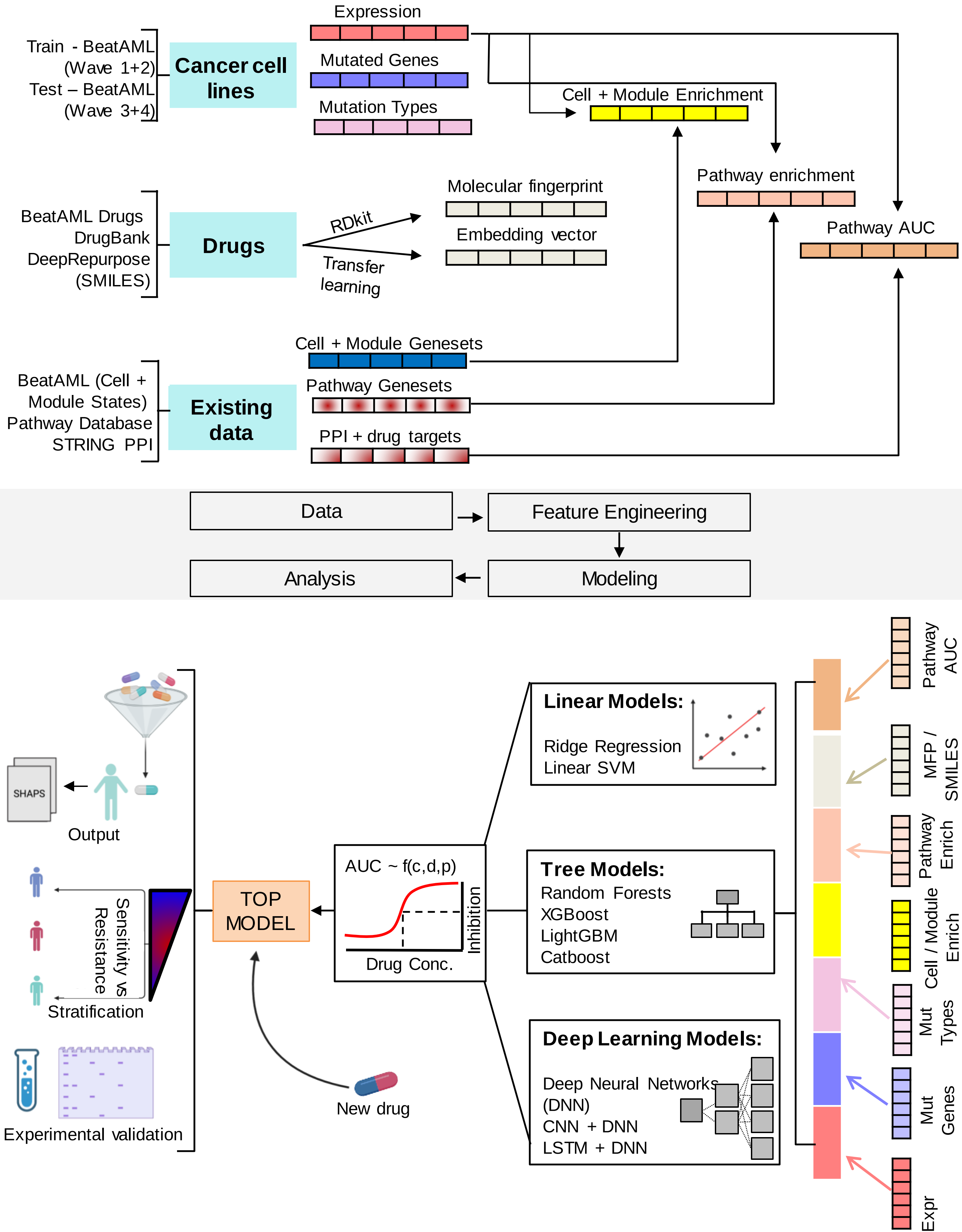
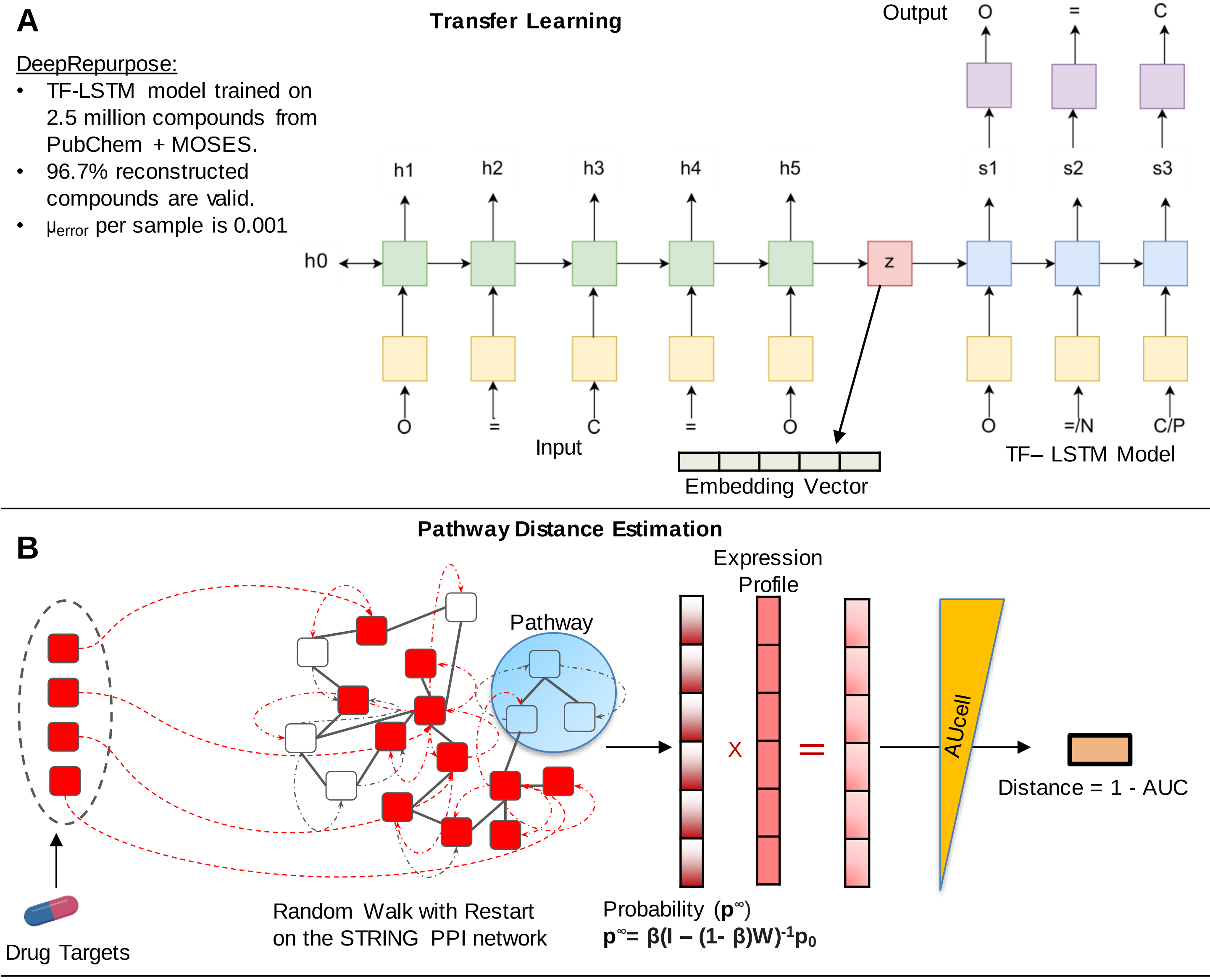

Train - BeatAML
(Wave 1+2)
Test – BeatAML
(Wave 3+4)
Cancer cell lines
Expression
Mutated Genes
Mutation Types
Cell + Module Enrichment
Pathway enrichment
Pathway AUC
BeatAML Drugs
DrugBank
DeepRepurpose
(SMILES)
Drugs
RDkit
Transfer learning
Molecular fingerprint
Embedding vector
BeatAML (Cell +
Module States)
Pathway Database
STRING PPI
Existing data
Cell + Module Genesets
Pathway Genesets
PPI + drug targets
Data
Feature Engineering
Modeling
Analysis
SHAPS
Output
Sensitivity vs
Resistance
Stratification
Experimental validation
TOP MODEL
New drug
AUC ~ f(c,d,p)
Inhibition
Drug Conc.
Linear Models:
Ridge Regression
Linear SVM
Tree Models:
Random Forests
XGBoost
LightGBM
Catboost
Deep Learning Models:
Deep Neural Networks
(DNN)
CNN + DNN
LSTM + DNN
Pathway AUC
MFP / SMILES
Pathway Enrich
Cell / Module Enrich
Mut Types
Mut Genes
Expr

Figure 2



DeepRepurpose:

- TF-LSTM model trained on 2.5 million compounds from PubChem + MOSES.
- 96.7% reconstructed compounds are valid.
- μ_{error} per sample is 0.001

Output
O
=
C
h1
h2
h3
h4
h5
s1
s2
s3
h0
z
O
=
C
=
O
O
=/N
C/P
Input
TF– LSTM Model
Embedding Vector
B
Pathway Distance Estimation
Expression Profile
Pathway
x
=
AUcell
Distance = 1 - AUC
Drug Targets
Random Walk with Restart on the STRING PPI network
Probability ($\mathbf{p}^{\infty}$)
$\mathbf{p}^{\infty} = \boldsymbol{\beta}(\mathbf{I} - (1 - \boldsymbol{\beta})\mathbf{W})^{-1}\mathbf{p}_0$

Figure 3

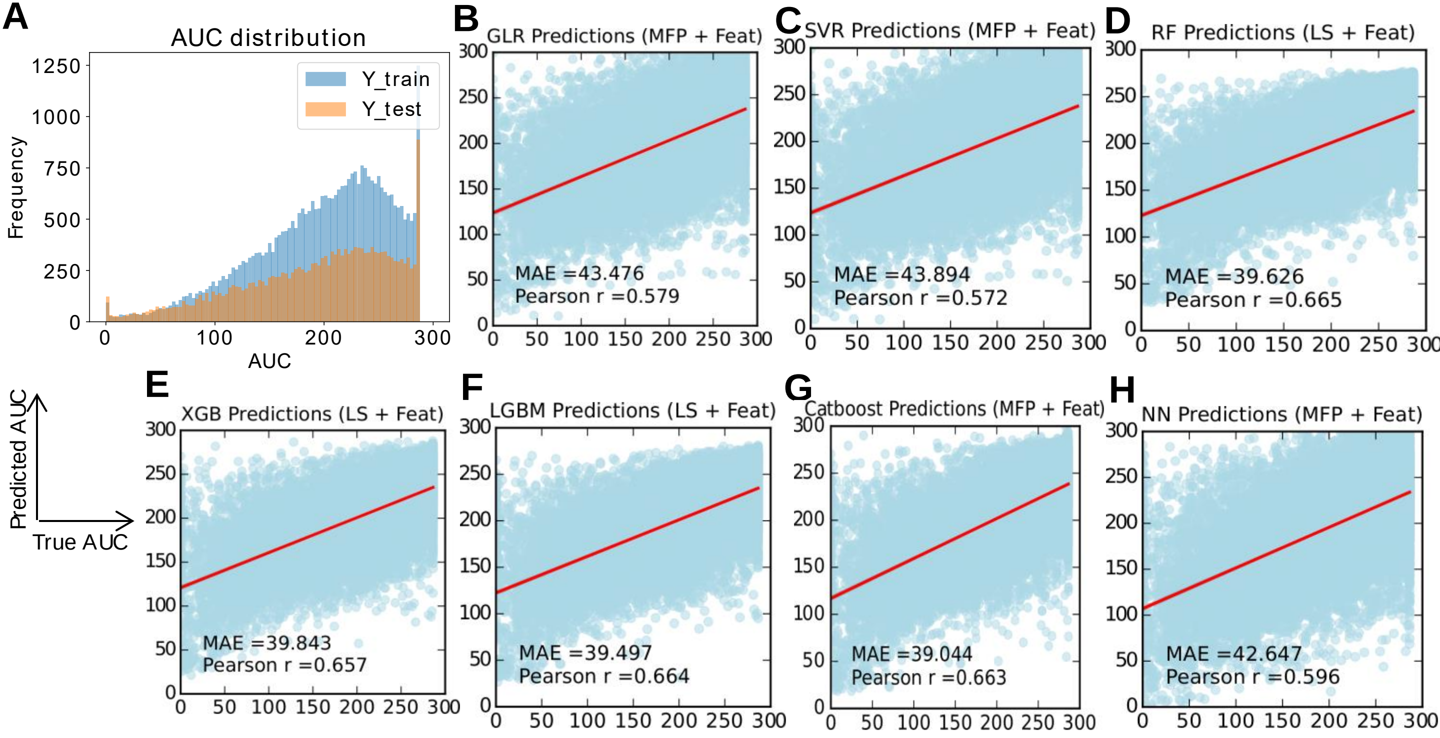
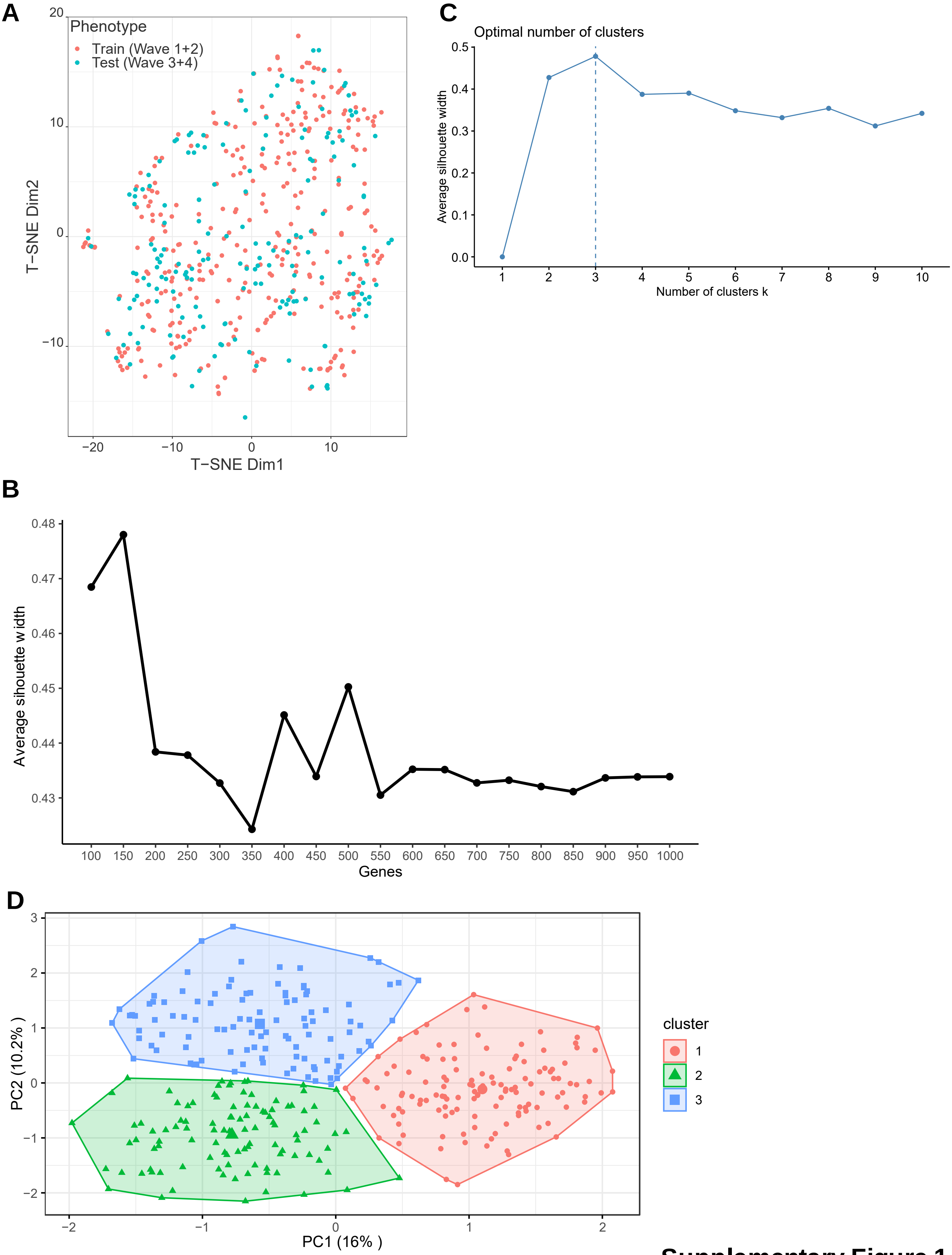
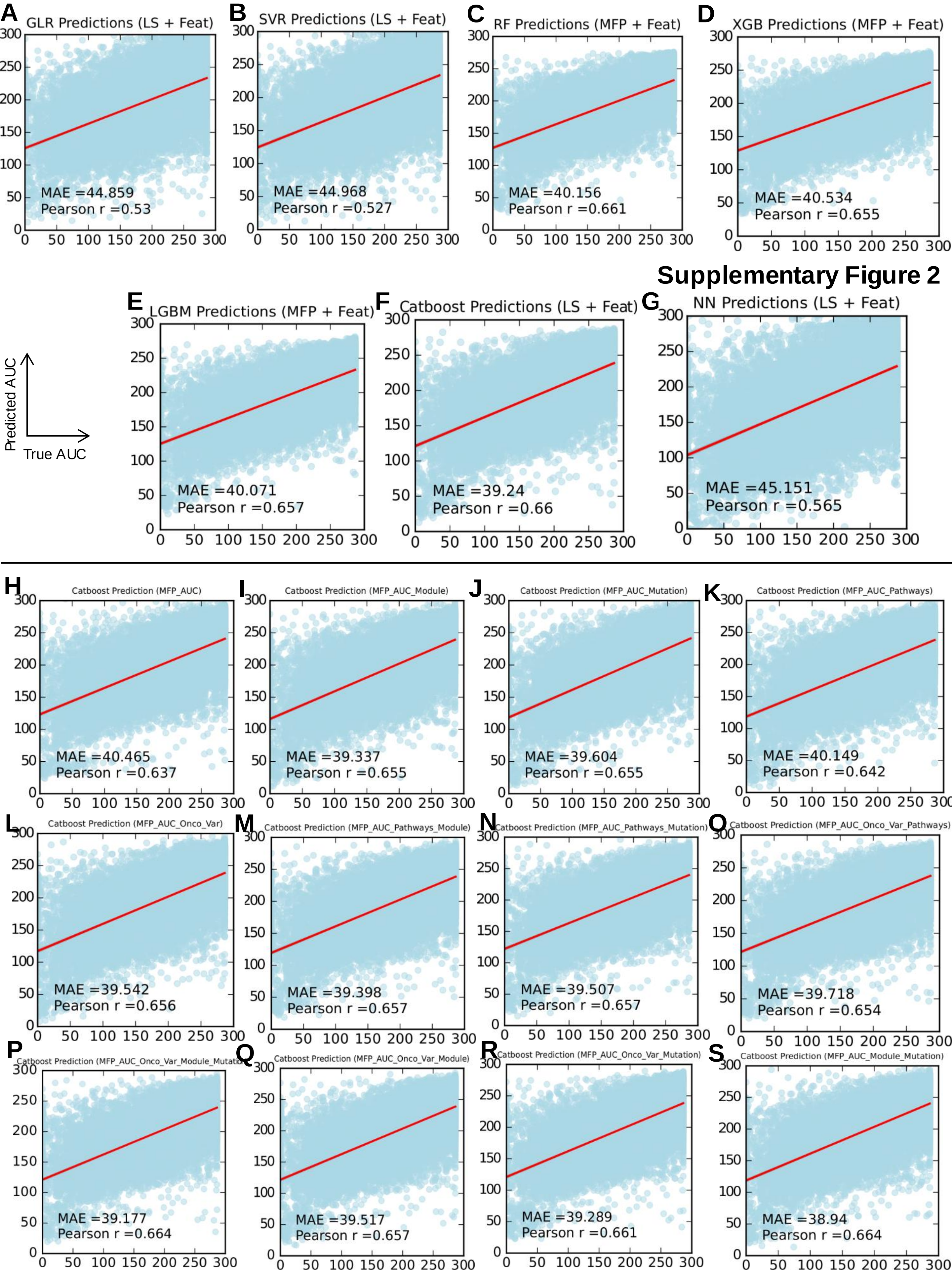

A
AUC distribution
Y_train
Y_test
Frequency
AUC
0
250
500
750
1000
1250
0
100
200
300
B
GLR Predictions (MFP + Feat)
MAE =43.476
Pearson r =0.579
C
SVR Predictions (MFP + Feat)
MAE =43.894
Pearson r =0.572
D
RF Predictions (LS + Feat)
MAE =39.626
Pearson r =0.665
Predicted AUC
True AUC
E
XGB Predictions (LS + Feat)
MAE =39.843
Pearson r =0.657
F
LGBM Predictions (LS + Feat)
MAE =39.497
Pearson r =0.664
G
Catboost Predictions (MFP + Feat)
MAE =39.044
Pearson r =0.663
H
NN Predictions (MFP + Feat)
MAE =42.647
Pearson r =0.596
0
50
100
150
200
250
300

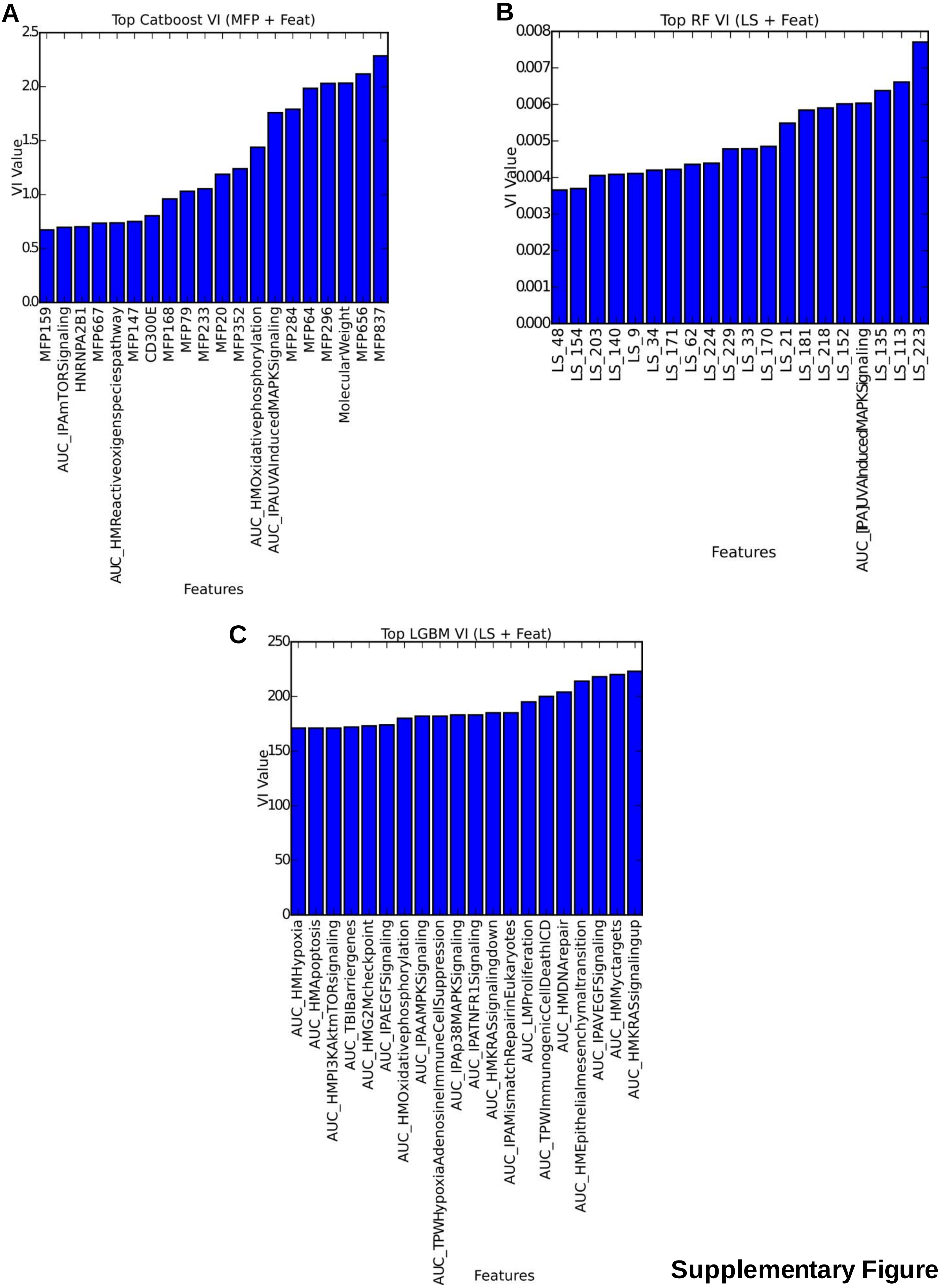
Figure 4



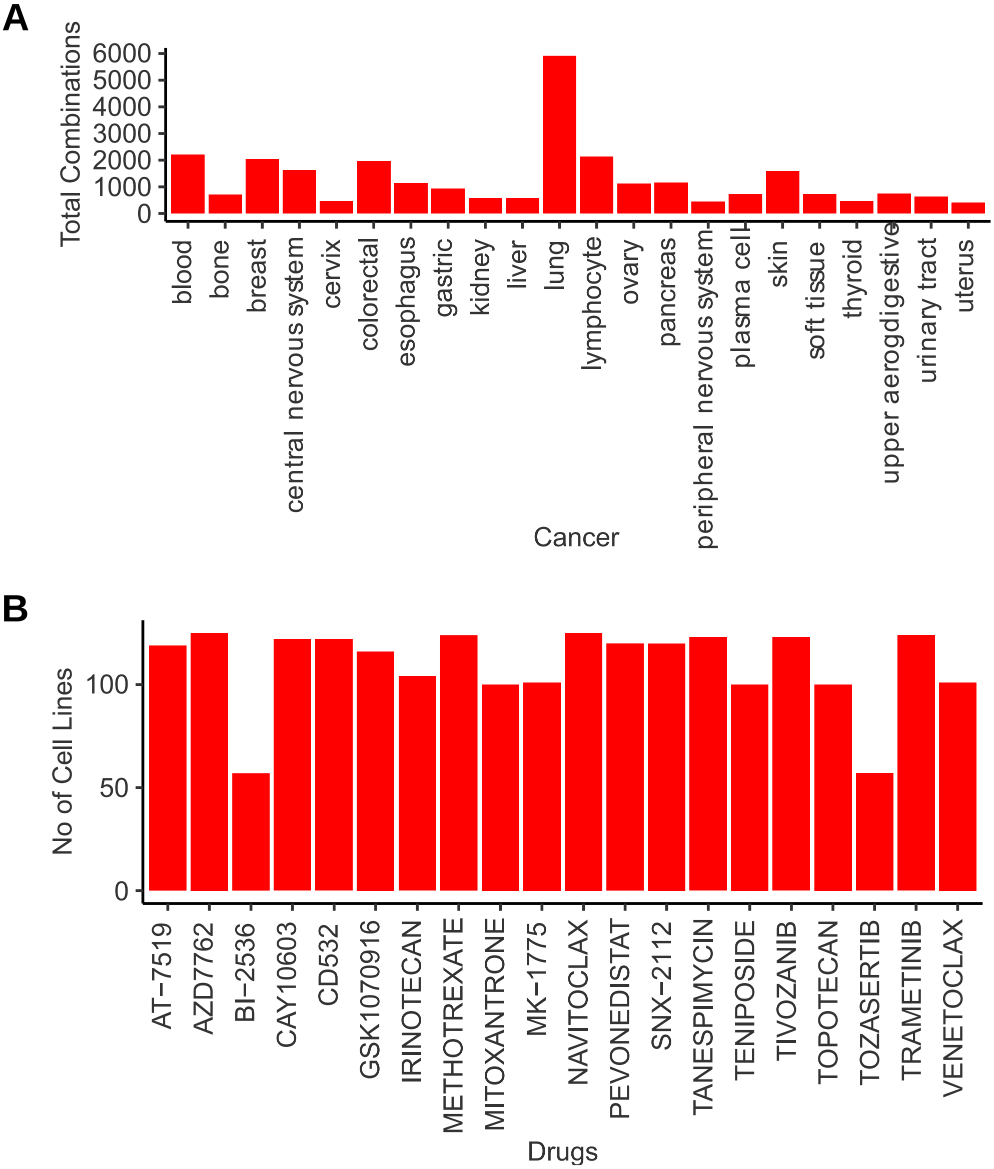
Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4

Methods	Features Used	CV (MAE)	Test (MAE)	CV (RMSE)	Test (RMSE)	CV (r2)	Test (r2)	CV Pearson (r)	Test Pearson (r)	CV Spearman (r)	Test Spearman (r)
Ridge Regressor	LS + Cell Line	30.224 +/- 1.972	44.859	40.342 +/- 2.906	60.005	0.555 +/- 0.052	0.281	0.744 +/- 0.034	0.53	0.742 +/- 0.032	0.556
Ridge Regressor	MFP + Cell Line	29.925 +/- 1.917	43.476	40.103 +/- 2.891	57.257	0.559 +/- 0.052	0.336	0.747 +/- 0.034	0.579	0.745 +/- 0.033	0.578
Linear SVR	LS + Cell Line	30.125 +/- 1.955	44.968	40.249 +/- 2.898	60.366	0.556 +/- 0.051	0.278	0.745 +/- 0.034	0.527	0.743 +/- 0.032	0.557
Linear SVR	MFP + Cell Line	29.934 +/- 1.922	43.894	40.088 +/- 2.894	57.731	0.559 +/- 0.052	0.328	0.747 +/- 0.034	0.572	0.745 +/- 0.033	0.571
RF	LS + Cell Line	**21.824 +/- 0.973**	**39.626**	**29.865 +/- 1.514**	**52.077**	**0.787 +/- 0.021**	**0.442**	**0.887 +/- 0.012**	**0.665**	**0.886 +/- 0.009**	**0.656**
RF	MFP + Cell Line	20.749 +/- 1.221	40.156	28.287 +/- 1.834	52.528	0.817 +/- 0.021	0.437	0.904 +/- 0.012	0.661	0.9 +/- 0.01	0.649
XGBoost	LS + Cell Line	24.288 +/- 1.212	39.843	31.939 +/- 1.755	52.339	0.728 +/- 0.038	0.432	0.853 +/- 0.022	0.657	0.835 +/- 0.022	0.65
XGBoost	MFP + Cell Line	22.937 +/- 1.009	40.534	30.101 +/- 1.445	52.876	0.782 +/- 0.037	0.43	0.884 +/- 0.21	0.655	0.867 +/- 0.022	0.642
LightGBM	LS + Cell Line	**21.529 +/- 1.119**	**39.497**	**28.748 +/- 1.63**	**52.091**	**0.786 +/- 0.029**	**0.441**	**0.887 +/- 0.016**	**0.664**	**0.872 +/- 0.016**	**0.655**
LightGBM	MFP + Cell Line	23.232 +/- 1.079	40.071	31.123 +/- 1.651	52.554	0.752 +/- 0.037	0.432	0.867 +/- 0.021	0.657	0.854 +/- 0.021	0.644
Catboost	LS + Cell Line	26.35 +/- 1.559	39.24	37.256 +/- 2.51	52.449	0.623 +/- 0.044	0.436	0.789 +/- 0.028	0.66	0.789 +/- 0.024	0.655
Catboost	MFP + Cell Line	**21.02 +/- 1.48**	**39.044**	**31.919 +/- 2.432**	**52.032**	**0.725 +/- 0.036**	**0.439**	**0.851 +/- 0.021**	**0.663**	**0.85 +/- 0.017**	**0.652**
DNN	LS + Cell Line	23.511 +/- 1.431	45.151	32.091 +/- 2.1	58.69	0.718 +/- 0.032	0.32	0.847 +/- 0.019	0.565	0.836 +/- 0.018	0.552
DNN	MFP + Cell Line	18.189 +/- 1.05	42.647	25.422 +/- 1.582	56.177	0.824 +/- 0.018	0.356	0.908 +/- 0.01	0.596	0.899 +/- 0.01	0.587
Graph Attention Network + FFNN	SMILES + Cell Line										
CNN + FFNN	SMILES + Cell Line										
LSTM + FFNN	SMILES + Cell Line										

Table 1: Performance comparison of different machine learning models

Features Used	CV (MAE)	Test (MAE)	CV (RMSE)	Test (RMSE)	CV (r2)	Test (r2)	CV (Pearson r)	Test (Pearson r)	CV (Spearman r)	Test (Spearman r)
MFP + AUC	27.458 +/- 1.207	40.465	38.85 +/- 1.856	54.114	0.591 +/- 0.054	0.406	0.768 +/- 0.035	0.637	0.774 +/- 0.028	0.626
MFP + Onco +Var + AUC	20.862 +/- 1.39	39.542	31.731 +/- 2.305	52.46	0.728 +/- 0.037	0.43	0.853 +/- 0.022	0.656	0.852 +/- 0.017	0.642
MFP + Pathways + AUC	19.556 +/- 1.005	40.149	30.409 +/- 1.866	53.287	0.751 +/- 0.03	0.412	0.866 +/- 0.017	0.642	0.867 +/- 0.014	0.628
MFP + Modules + AUC	20.996 +/- 1.183	39.337	31.81 +/- 2.001	52.509	0.728 +/- 0.033	0.429	0.853 +/- 0.019	0.655	0.854 +/- 0.015	0.644
MFP + Mutations + AUC	22.004 +/- 1.185	39.604	32.911 +/- 1.944	52.854	0.709 +/- 0.038	0.429	0.842 +/- 0.022	0.655	0.843 +/- 0.018	0.64
MFP + AUC + Pathways + Modules	25.35 +/- 1.324	39.398	36.136 +/- 2.137	52.482	0.646 +/- 0.044	0.432	0.803 +/- 0.027	0.657	0.804 +/- 0.023	0.647
MFP + AUC + Pathways + Mutations	25.59 +/- 1.377	39.507	36.397 +/- 2.138	52.792	0.641 +/- 0.044	0.431	0.8 +/- 0.027	0.657	0.801 +/- 0023	0.648
MFP + AUC + Pathways + Onco +Var	24.765 +/- 1.473	39.718	35.574 +/- 2.351	52.744	0.658 +/- 0.042	0.427	0.81 +/- 0.026	0.654	0.81 +/- 0.021	0.643
MFP + AUC + Modules + Mutations	**25.858 +/- 1.415**	**38.94**	**36.688 +/- 2.168**	**52.229**	**0.635 +/- 0.045**	**0.441**	**0.796 +/- 0.028**	**0.664**	**0.797 +/- 0.023**	**0.656**
MFP + AUC + Modules + Onco + Var	24.768 +/- 1.405	39.517	35.575 +/- 2.239	52.653	0.657 +/- 0.042	0.432	0.81 +/- 0.026	0.657	0.809 +/- 0.021	0.648
MFP + AUC + Mutations + Onco + Var	24.749 +/- 1.477	39.289	35.578 +/- 2.396	52.387	0.657 +/- 0.044	0.437	0.81 +/- 0.027	0.661	0.81 +/- 0.022	0.651
MFP + AUC + Mutations + Modules + Onco + Var	**24.743 +/- 1.51**	**39.177**	**35.559 +/- 2.445**	**52.323**	**0.657 +/- 0.043**	**0.44**	**0.81 +/- 0.026**	**0.664**	**0.81 +/- 0.021**	**0.655**
MFP + AUC + Mutations + Pathways + Onco + Var	25.675 +/- 1.556	39.803	36.518 +/- 2.458	53.268	0.638 +/- 0.045	0.419	0.798 +/- 0.028	0.647	0.797 +/- 0.023	0.638
MFP + AUC + Modules + Pathways + Onco + Var	24.718 +/- 1.461	39.422	35.531 +/- 2.354	52.581	0.658 +/- 0.046	0.433	0.811 +/- 0.028	0.658	0.81 +/- 0.023	0.648
MFP + AUC + Mutations + Modules + Pathways	27.653 +/- 1.425	39.403	38.459 +/- 2.177	52.601	0.599 +/- 0.48	0.436	0.774 +/- 0.03	0.66	0.775 +/- 0.026	0.652

Supplementary Table 1: **Ablation study of different feature sets used for the optimal model construction.**

Background:

1. 699 cell lines from CCLE with 19,177 gene expression profiles. This information was downloaded from Cancer Dependency Map Public 21Q3. We filtered the original data consisting of 1,377 cell lines → 699 cell lines to keep only those cell lines with COSMIC ids to match the drug-response dataset from GDSC portal. We include 10 features related to cell line metadata including age, gender, type, name of cell line etc. for the cancer cell lines.

2. Genes of interest include genes which are part of several inflammasome/inflammatory cell death pathways including:
a) Reactome inflammasome, b) KEGG nod like signaling pathway, c) GO biological process inflammasome complex, d) Reactome pyroptosis, e) Necroptotic signaling pathway from GO, f) PANoptosis pathway, g) Immunogenic cell death pathway (ICR) → total of 170 (167 of which are present in the 19,177 genes)

3. Seven pathways considered for inflammatory cell death as mentioned above.

4. We got the mutation profile and copy number variation profile for the 170 genes of interest. The mutation profile and copy number variation was obtained from Harmonizome database from Mayan lab.

5. We removed genes which had no variation in expression, mutation or CNV across cell lines including: Mutation_ERBIN, Mutation_NLRP2B, Mutation_STMP1, Mutation_PYDC2, Mutation_CARD18, CNV_ERBIN, CNV_NLRP2B, CNV_STMP1

6. Total Features include:
a) Cell Line Features (10); b) Pathways (7); c) Expression (167); d) Mutation (162); e) CNV (164)

7. The dose response information is obtained from GDSC portal. It contains drug response for a particular cell line with prediction variables: IC50score and Z-score. In the GDSC portal, the Z-score is used to determine sensitive and resistant drugs with cut-offs of -2 and 2 respectively. We use the –log10(IC50score) as our y variable (term to predict). It contains 398 unique drugs and 989 cell lines.

8. The viability information is also obtained from GDSC portal. It contains cell viability at different dosage levels. **Currently not used.**

Methods:

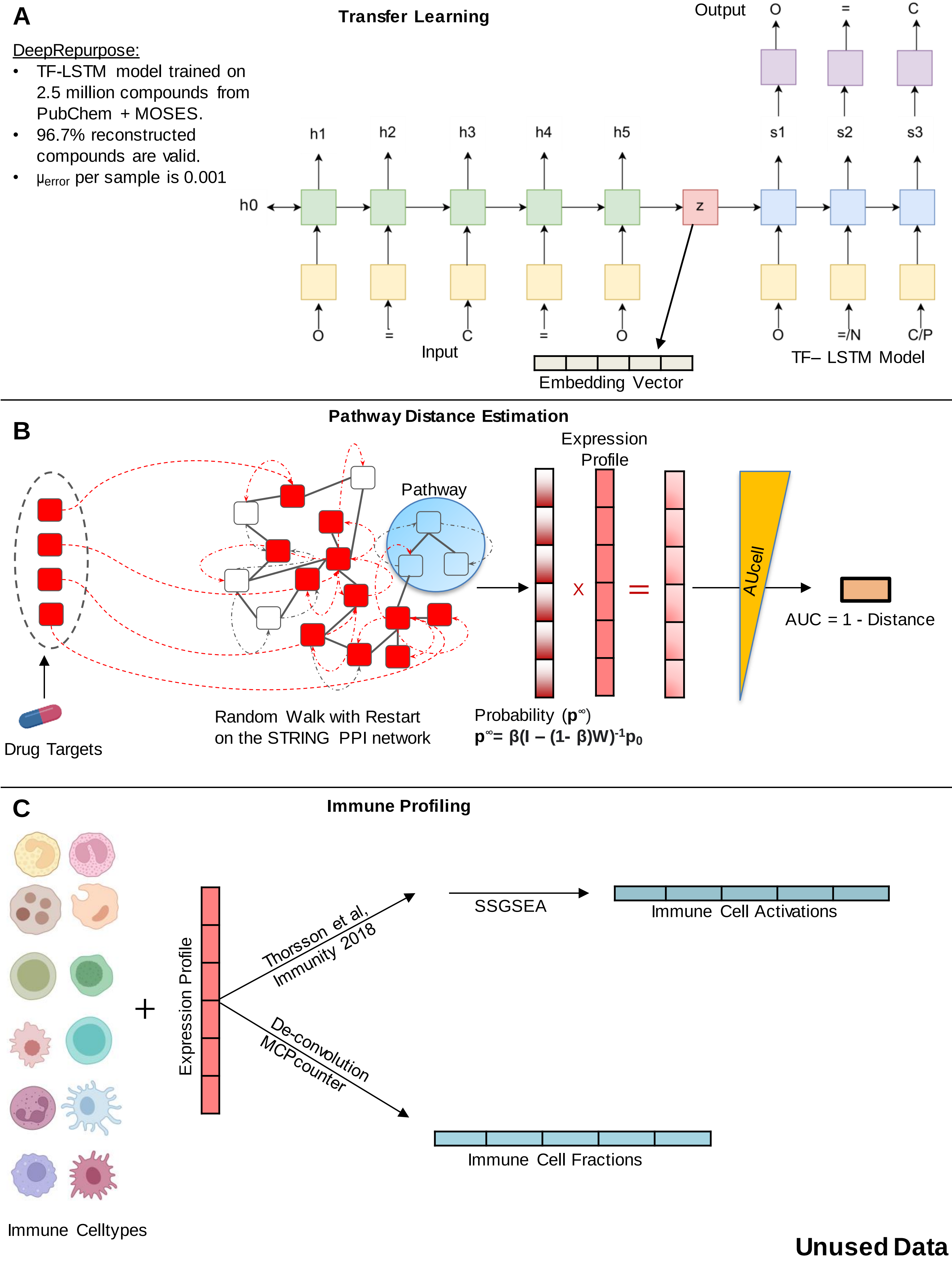
1. The gene expression profile for each cell line is quantile normalized (19,177 genes), scaled and converted to z-scores.
2. T-sne plots are made for 699 cell lines with cancer types based on expression profiles as well as the features we have engineered.
3. A complexheatmap visualization of the different features used and how they look across the 699 cancer cell lines.

Background:

1. We obtained drug information for 399 drugs from GDSC portal and the cell line information for 699 cell lines from CCLE.
2. We curated and added each drug's target from resources like Drug Bank, its SMILES, Inchikey, Molecular weight, molecular formula representations from PubChem, CheMBL and NCBI API.
3. There are a total of 373 unique drugs available from this dataset.
4. We find a total of 379,005 drug-cell line sensitivity profiles from GDSC portal (GDSC 1 & 2 combined).
5. After filtering, removing duplicates and combined drug plus cell line features, we end up with a total of 151,636 drug-cell line combination profiles consisting of 253 drugs and 693 cell lines.

Method:

1. We perform inner joins with drug profiles and cell line profiles to get drug-cell features used for the predictive models.
2. For each drug, we have its known targets. We perform a random walk with restart to get a random walk score for each drug using 'diffusr' package in R. It provides the probability of a drug impacting all the genes in our cell line's expression profile. This affinity is based on topology and doesn't consider individual gene's expression in a particular cell line. To get a cell line specific affinity, we multiply the random walk scores with corresponding genes' expression levels. Using this information and the gene set for each pathway, we estimate the distance of a drug from each of the 7 inflammasome related pathways using 'AUCell' package in R.
3. For each drug, we can estimate its molecular fingerprint representation using the RDKIT package in python.
4. For each drug, we can also estimate its representation using a transfer-learning based approach. We pass the drug SMILES to a TF-LSTM autoencoder trained on over 2 million drug SMILES and obtain a embedding vector representation of the drug which can be fed to a machine learning algorithm.
5. All the cell line information (expression, mutation, copy number) + drug information (vector representation) + pathway enrichments (pathway activation based on expression) and distance of a pathway from a drug's known targets are used as features to predict the drug response.
6. We used a variety of machine learning methods including:
    1. Linear Regression
    2. Elastic Net
    3. SVM
    4. Random Forests
    5. Xgboost
    6. LightGBM
    7. Feed Forward Neural Networks (DNN)
    8. Graph Attention Network (GAT) + DNN
    9. Convolutional Neural Network (CNN) + DNN
    10. Long-Short Term Memory (LSTM) + DNN
7. We built different models on training set (565 cell lines) and test on a complete independent test set (128 cell lines on which no training is performed)
8. We highlight the variable importance (i.e. the top features) driving the prediction for each of these different machine learning models and performance is highlighted in Table.



DeepRepurpose:

- TF-LSTM model trained on 2.5 million compounds from PubChem + MOSES.
- 96.7% reconstructed compounds are valid.
- μ_{error} per sample is 0.001

Research Article Checklist should be placed after presentation

Research Article Checklist

First author and coauthors are responsible for ensuring that: 1) all statements are factually accurate; 2) references to seminal publications and Kanneganti lab publications are used consistently and wherever appropriate (refer to the reference lists); 3) the correct mouse source and information is provided.

- □ **Title**: informative, concise, and includes keywords
- □ **Author List**: all authors are included; names are spelled correctly
- □ **Keywords**: searchability; citability
- □ **Abstract**: Overall summary (provides key background, purpose, and findings)
  - ○ Statements are factually accurate
  - ○ Concise
  - ○ Followed example
- □ **Outline or Full Article**:
  - ○ Introduction: 3 paragraphs; key concepts and background
  - ○ Statements are factually accurate
  - ○ **References** (Key refs from Kanneganti lab and Others ref lists)
  - ○ Discussion: 3 paragraphs; concise summary of info presented; contextualized; highlights any clinical relevance
  - ○ Followed examples
- □ **Figures**:
  - ○ Concise title and legend
  - ○ Labels and enough detail to be understood without text
  - ○ Followed lab template (correct formatting, no shadows, no typos)
  - ○ No image duplications

I confirm that I have checked for the above points during my review and provided the necessary feedback to the first author.

Name Date